PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G10L 15/22 — A1

(11) International Publication Number: **WO 00/36591**

(43) International Publication Date: 22 June 2000 (22.06.00)

(21) International Application Number: PCT/EP99/09263

(22) International Filing Date: 29 November 1999 (29.11.99)

(30) Priority Data:
98204286.3 17 December 1998 (17.12.98) EP

(71) Applicant: KONINKLIJKE PHILIPS ELECTRONICS N.V. [NL/NL]; Groenewoudseweg 1, NL–5621 BA Eindhoven (NL).

(71) Applicant *(for DE only)*: PHILIPS CORPORATE INTELLECTUAL PROPERTY GMBH [DE/DE]; Habsburgerallee 11, D–52066 Aachen (DE).

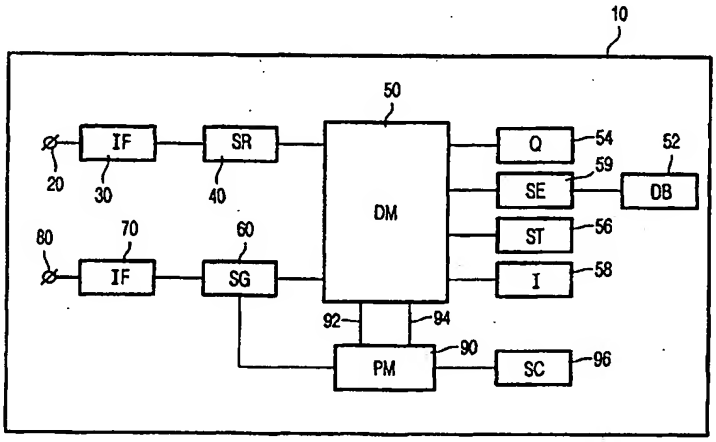
(72) Inventor: PANKERT, Matthias; Prof. Holstlaan 6, NL–5656 AA Eindhoven (NL).

(74) Agent: HOEKSTRA, Jelle; Internationaal Octrooibureau B.V., Prof. Holstlaan 6, NL–5656 AA Eindhoven (NL).

(81) Designated States: JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: SPEECH OPERATED AUTOMATIC INQUIRY SYSTEM



(57) Abstract

An automatic inquiry system includes a storage (52) for storing a plurality of inquiry entries, each entry comprising a plurality of data fields. For at least one data field of each inquiry entry a respective reference transcription set is stored which includes at least two alternative acoustic transcriptions of data stored in the data field. A dialogue engine (50) executes a human–machine dialogue to determine a plurality of pre–determined query fields. Instead of fully recognizing all utterances, a speech recognizer (40) is able to represent at least one utterance as a corresponding input transcription set, which includes at least two alternative acoustic transcriptions of the utterance. The dialogue engine (50) specifies at least a first one of the query fields by an input transcription set of a selected one of the utterances. A search engine (59) locates inquiry entries in the storage whose associated reference transcription set relates to the first one of the query fields. In this way, utterances, such as specifying family names, which are difficult to recognize accurately need not be fully recognized before a query is made.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

Speech operated automatic inquiry system.

The invention relates to an automatic inquiry system including:

a storage for storing a plurality of inquiry entries, each inquiry entry comprising a plurality of data fields;

a dialogue engine for executing a machine-controlled human-machine dialogue to determine a plurality of pre-determined query fields;

a speech recognizer operative to represent an utterance as a corresponding input transcription set, which includes at least two alternative acoustic transcriptions of the corresponding utterance; a sequence of utterances specifying the respective query fields; and

a search engine for locating inquiry entries in the storage in dependence on the query fields.

Automatic inquiry systems, for instance for obtaining directory information, increasingly use automatic human-machine dialogues. Typically, a user establishes a connection with the inquiry system using a telephone. In a dialogue between the machine and the user, a dialogue engine tries to establish a number of query items enabling obtaining the desired information from a storage, such as a database. The query items specify the data of corresponding query fields. During the dialogue, the system typically issues an initial question. For a directory assistance system, such an initial question might be "Whose telephone number would you like to have ?". During the following dialogue, the system uses speech recognition techniques to extract the query items from the user's utterances. For example, a directory assistance system needs to obtain at least a few of the following items to perform a query: the family name, given name, town, street and street number of a person (or similar items for a business). The dialogue may be follow a step-wise approach where the system prompts the user for one query item at a time and the system assumes that information recognized from the response specifies the query item. For instance, for a directory assistance system the system could successively ask the user to specify the town, family name, and street. As an alternative, the system may allow a free-speech dialogue where the system extracts the relevant query items from the sequence of utterances of the user. Once all essential query items have been

recognized, the answer(s) to the query are obtained from the storage. This results in one or more information items. The information items are typically presented to the user in spoken form. Normally, textual presentations of the information items are inserted in a sequence of preformatted phrase or sentence templates. Usually, the templates are prosodically enriched. If no suitable speech presentation is available, for instance in the form of sampled phrases/words, the prosodically enriched text may be converted to speech using a speech synthesis technique. The speech output is fed back to the user. As an alternative to presenting information to the user, the system may also act on the data. For instance, a directory assistance system could automatically connect the user to the retrieved telephone number.

In known automatic inquiry systems, the speech recognizer as a first step in the recognition represents utterances of the user as a set of acoustic transcriptions. Typically, an acoustic transcription represents a time-sequence of phonemes or similar subword units. The set includes several alternative transcriptions since normally several sequences of phonemes/sub-word units are acoustically close to the utterance. Frequently, Hidden Markov Models are used to identify several likely acoustic transcriptions. In a next step, the speech recognizer makes a selection between the various transcriptions. To this end, the transcriptions are converted into textual representations. The textual representations are compared to a lexicon which comprises all textual representations allowed for the query item in question. For instance, for the query field "town" the lexicon comprises all names of towns in the area for which the system is operational. Based on this comparison constraint, the number of textual representations is reduced. The dialogue engine is then provided with a limited number of textual representations for each query field. The dialogue engine performs a query on the storage to determine one or more inquiry entries which match all or a selected few of the query items.

Using the technique described above for large inquiry systems, and in particular for large directory assistance systems, is cumbersome. Whereas the known technique operates satisfactorily for systems with relatively small lexicons (e.g. in the order of tens of thousand entries in the lexicon), the recognition result of query items may drop to 20% or even lower for lexicons with a million or more entries. Such large lexicons are particularly required to represent family names in nation-wide telephone directory assistance systems. Particularly foreign names may be pronounced in many different ways, resulting in a much larger variation in possible transcriptions associated with such a name. In fact, the variation is much larger than a normal variations where the actual pronunciations sound very similar to a preferred transcription. For names an actual pronunciation may sound entirely different from a preferred

transcription. As an example, a Celtic pronunciation of an Irish name usually bares hardly any resemblance to an Anglo-Saxon pronunciation of the same name. Moreover, for free-speech dialogues it is not known up front which item is being specified by the user. As a consequence, each utterance has to be compared against the full lexicon. Inquiry systems where one of the query fields specifies a family name furthermore have to cope with the fact that the spread in pronunciation is larger than for normal words. Moreover, the conversion of an acoustical transcription to a textual form is difficult, since in many cases standard rules for converting a phonetic transcription to text will not produce the desired result. As an example, the Dutch names "Jansen" and "Janssen" are pronounced in the same way but have a different textual representation. Similarly, a same pronunciation may be used for "Martineau", "Martinho", "Martinow", and "Martinau".

It is an object of the invention to provide an automatic inquiry system of the kind set forth which is capable of overcoming at least some of above mentioned problems. In particular it is an object to provide such a system which is capable of coping with query items, such as representing names, with a large spread in textual representation for a similar pronunciation and/or a large spread in pronunciation for a same textual representation.

To meet the object of the invention, the system is characterized in that <refer characterizing part of claim 1>. According to the invention, the searching of the storage is not performed on textual representations of the utterances but on acoustic representations. At least one of the query fields is specified as a set of acoustic transcriptions and compared to a set of reference acoustic transcriptions in the storage. Other query fields may still be represented in a textual form. By storing a reference acoustic transcription, the speech recognizer is relieved from performing the difficult process of representing the utterance in a textual form. By storing alternative acoustic transcriptions, the system can cope in an easy way with a wide spread in pronunciation.

It is known from the National Directory Inquiry System (NDIS) of pc-plus INFORMATIK GmbH in Munich, Germany, to enter query items via the keypad of a telephone. To cope with the fact that entries may sound similar but are spelled differently (i.e. the user has spelled a name different than the desired entry in the database), phonetic conversion rules are used to convert the textual entry into alternative textual entries which sound similar. This system does not support speech-input but instead requires input of characters via a numeric keypad, which is particularly difficult in most European countries where characters are not indicated on the keypad. Moreover, a user who in many cases does

not know a correct spelling of a name, but does know a correct pronunciation, must 'choose' a spelling. The system then, starting from the user-provided spelling, selects alternative, similar sounding spellings. If the user made an unfortunate spelling (which may be caused by entering only one wrong character), the system may interpret the entry as sounding different than intended and not being able to find the desired alternative textual form. Consequently, the system will not be able to locate the desired inquiry entry. By using the knowledge of the user of pronouncing entries, phonetic conversion routines, which inevitably reduce the success rate, are avoided. For difficult entries, such as foreign names, where the spread in pronunciation is larger than normal, many alternative reference acoustic transcriptions may be stored. It will be appreciated that according to the invention, a correct textual transcription of the entry is not required for the searching. The storage may still comprise such a textual transcription for supply to the user or other purposes.

In an embodiment as defined in the dependent claim 2, at least two query items are represented by transcription sets. A combined search is performed on those (and possibly other) query items. In the conventional dialogue system, the entries were individually recognized in the sense that a few most likely candidates were output for each entry. Afterwards, a combined search was performed for the remaining candidates. In a practical situation, a specific family name may sound similar to hundreds of names, e.g. two hundred names. The number of candidates is then reduced to, for instance, ten most likely candidates. In order to make this reduction, millions of family names may need to be compared to the acoustic transcription of the utterance. A specific town may sound similar to a few dozen names, e.g. twenty town names, also being reduced to ten most likely candidates at the cost of many comparisons (for free-speech dialogues this may also involve millions of comparisons). The database is then used to verify whether any of the possible 100 remaining options actually exists. In the reduction process the actual correct entries may have been filtered out. In fact, for the total number of 4000 (200x20) similar sounding entries, the database may only have one or very few matches. By using the database to perform a combined matching operation, no extensive pre-filtering is required. Any suitable database technique may be used to achieve a fast searching. For instance, indexing may be used on entries with least candidates. In a directory assistance system, advantageously indexing on a town name is used, reducing the search space for family names. By also indexing on street names, the search space for family names can in most cases be reduced to only a few dozen or to a few hundred.

In an embodiment as defined in the dependent claim 3, two sets of acoustic transcriptions are considered related (i.e. matching) if at least one transcription of the first set

is acoustically similar to at least one transcription of the second set. It is not required that all transcriptions of the sets are acoustically similar. This allows for a wide variation in pronunciation. Any conventional technique for determining that two transcriptions are acoustically similar may be used, for instance by using similarity measures or a similarity ranking which indicate similarity of two phonemes, for each phoneme pair.

In an embodiment as defined in the dependent claims 4, 5 and 6, at least one set of transcriptions is represented as a graph. This allows for an effective representation where paths through the graph may diverge, where part of a 'word' is pronounced differently and paths may converge, where parts are pronounced similarly. In this way, a sharing of representation of the transcriptions occurs. An individual transcription is then represented by a path through the graph. Preferably, the comparison of sets is not performed by isolating all individual transcriptions and comparing the individual transcriptions, but by using the sharing given the graph. As an example, if a particular path through a graph of the first set at a certain position does not match a path through the graph of the second set (or an individual transcription of the second set), then all transcriptions coinciding with the beginning of the particular path (until the mis-match position) will also not match and do not need to be compared separately.

In an embodiment as defined in the dependent claim 7, the set of acoustic transcriptions is pruned to reduce the number of comparisons to be made. Since in the system according to the invention no full recognition of an utterance is required, this enables further pruning than in conventional systems.

In an embodiment as defined in the dependent claim 8, the pruning is based on acoustic similarity. For instance, the reference set of acoustic transcriptions could be reduced by only keeping transcriptions representing very different pronunciations, for instance to cover entirely different name pronunciations. Transcriptions representing very similar pronunciations can be reduced to one or a few representing the most common ('average') pronunciation. The input transcription set may be reduced in a conventional manner to represent the acoustically most likely transcription.

In an embodiment as defined in the dependent claim 9, a statistical model, which for instance specifies the likelihood of phoneme sequences, is used to eliminate unlikely acoustical transcriptions.

In an embodiment as defined in the dependent claim 10, a predetermined portion of the transcription is selected. This is possible in the system according to the invention, since no textual representation needs to be made. For instance, if an utterance

represents a family name it will be in most cases sufficient to only use the first three to five phonemes to be able to identify a directory entry, particularly in combination with other data such as street and town.

These and other aspects of the invention will be apparent from and elucidated with reference to the embodiments shown in the drawings.

Fig.1 shows a block diagram of the system according to the invention,

Fig. 2 shows a block diagram of the speech recognizer according to the invention,

Fig. 3 illustrates full word and sub-word models, and

Fig. 4 shows a database structure.

Figure 1 shows a block diagram of a system 10 according to the invention. Examples of the working of the system will be described in particular for a directory assistance system which automatically provides information with respect to telephone numbers (similar to the so-called white or yellow papers). It will be appreciated that these examples are not limiting. The system may equally well be used for supplying information like journey scheduling information, such as involving a train, bus or plane. Furthermore, the system may be used to supply other types of information, such as bank related information (e.g. an account overview), information from the public utility boards, information from the council or other governmental organizations, or, more in general, information related to a company (e.g. product or service information). The system may also act on the information, e.g. by making a phone call, instead of or in addition to presenting the information.

In figure 1, item 20 represents an interconnection for receiving a speech representative signal from a user. For instance, a microphone may be connected to the interconnection 20. Typically such a microphone is integrated in a telephone, allowing the user to operate the system remotely. The system comprises an interface 30 to receive the input from the user. This may for instance be implemented using a conventional modem. If the interface has an input for receiving speech in an analogue form, the interface preferably comprises an A/D converter for converting the analogue speech to digital samples of a format suitable for further processing by a speech recognition system 40. If the interface has an input for receiving the speech in a digital form, e.g. via ISDN, preferably the converter is capable of converting the digital data to a suitable digital format for further processing. Block 40

represents a speech recognition subsystem. The speech recognition system 40 typically analyses the received speech by comparing it to trained material (acoustical data and a lexicon/grammar). The speech recognition is preferably speaker-independent and allows continuous speech input. By itself, speech recognition is known and has been disclosed in various documents, such as EP 92202782.6, corresponding to US Serial No. 08/425,304 (PHD 91136), EP 92202783.4, corresponding to US Serial No. 08/751,377 (PHD 91138), EP 94200475.5, corresponding to US 5,634,083 (PHD 93034), all to the assignee of the present application. According to the invention at least one utterance of the user is not fully recognized (i.e. is not transcribed to a textual representation) by the speech recognition system 40. Instead, the speech recognizer 40 outputs a set of acoustic transcriptions representing the utterance. As will be described in more detail below, for many applications it is preferred that the speech recognizer 40 recognizes part of the sequence of utterances. With reference to Figs. 2 and 3, a description is given of an exemplary speech recognition system 40 which can fully recognize an utterance as well as represent the utterance as a set of acoustic transcriptions.

Speech recognition systems, such as large vocabulary continuous speech recognition systems, typically use a collection of recognition models to recognize an input pattern. For instance, an acoustic model and a vocabulary may be used to recognize words and a language model may be used to improve the basic recognition result. Figure 2 illustrates a typical structure of a large vocabulary continuous speech recognition system 40 [refer L.Rabiner, B-H. Juang, "Fundamentals of speech recognition", Prentice Hall 1993, pages 434 to 454]. The system 40 comprises a spectral analysis subsystem 210 and a unit matching subsystem 220. In the spectral analysis subsystem 210 the speech input signal (SIS) is spectrally and/or temporally analyzed to calculate a representative vector of features (observation vector, OV). Typically, the speech signal is digitized (e.g. sampled at a rate of 6.67 kHz.) and pre-processed, for instance by applying pre-emphasis. Consecutive samples are grouped (blocked) into frames, corresponding to, for instance, 32 msec. of speech signal. Successive frames partially overlap, for instance, 16 msec. Often the Linear Predictive Coding (LPC) spectral analysis method is used to calculate for each frame a representative vector of features (observation vector). The feature vector may, for instance, have 24, 32 or 63 components. The standard approach to large vocabulary continuous speech recognition is to assume a probabilistic model of speech production, whereby a specified word sequence $W = w_1w_2w_3...w_q$ produces a sequence of acoustic observation vectors $Y = y_1y_2y_3...y_T$. The recognition error can be statistically minimized by determining the sequence of words $w_1w_2w_3...w_q$ which most probably caused the observed sequence of observation vectors

$y_1y_2y_3...y_T$ (over time t=1, ... , T), where the observation vectors are the outcome of the spectral analysis subsystem 210. This results in determining the maximum a posteriori probability:

max P(W|Y), for all possible word sequences W

By applying Bayes' theorem on conditional probabilities, P(W|Y) is given by:

$$P(W|Y) = P(Y|W).P(W)/P(Y)$$

Since P(Y) is independent of W, the most probable word sequence is given by:

$$\arg\max P(Y \mid W).P(W) \text{ for all possible word sequences } W \qquad (1)$$

In the unit matching subsystem 220, an acoustic model provides the first term of equation (1). The acoustic model is used to estimate the probability P(Y|W) of a sequence of observation vectors Y for a given word string W. For a large vocabulary system, this is usually performed by matching the observation vectors against an inventory of speech recognition units. A speech recognition unit is represented by a sequence of acoustic references. Various forms of speech recognition units may be used. As an example, a whole word or even a group of words may be represented by one speech recognition unit. A word model (WM) provides for each word of a given vocabulary a transcription in a sequence of acoustic references. For systems, wherein a whole word is represented by a speech recognition unit, a direct relationship exists between the word model and the speech recognition unit. Other systems, in particular large vocabulary systems, may use for the speech recognition unit linguistically based sub-word units, such as phones, diphones or syllables, as well as derivative units, such as fenenes and fenones. For such systems, a word model is given by a lexicon 234, describing the sequence of sub-word units relating to a word of the vocabulary, and the sub-word models 232, describing sequences of acoustic references of the involved speech recognition unit. A word model composer 236 composes the word model based on the subword model 232 and the lexicon 234. Figure 3A illustrates a word model 300 for a system based on whole-word speech recognition units, where the speech recognition unit of the shown word is modeled using a sequence of ten acoustic references (301 to 310). Figure 3B illustrates a word model 320 for a system based on sub-word units, where the shown word is modeled by a sequence of three sub-word models (350, 360 and 370), each with a sequence of four acoustic references (351, 352, 353, 354; 361 to 364; 371 to 374). The word models shown in Fig. 3 are based on Hidden Markov Models (HMMs), which are widely used to stochastically model speech signals. Using this model, each recognition unit (word model or subword model) is typically characterized by an HMM, whose parameters are estimated from a training set of data. For large vocabulary speech recognition systems usually a limited set of, for instance 40, sub-word

units is used, since it would require a lot of training data to adequately train an HMM for larger units. An HMM state corresponds to an acoustic reference. Various techniques are known for modeling a reference, including discrete or continuous probability densities. Each sequence of acoustic references which relate to one specific utterance is also referred as an acoustic transcription of the utterance. It will be appreciated that if other recognition techniques than HMMs are used, details of the acoustic transcription will be different.

A word level matching system 230 of Fig. 2 matches the observation vectors against all sequences of speech recognition units and provides the likelihoods of a match between the vector and a sequence. If sub-word units are used, constraints can be placed on the matching by using the lexicon 234 to limit the possible sequence of sub-word units to sequences in the lexicon 234. This reduces the outcome to possible sequences of words. If for utterances, like representing family names, no full recognition is required, the involved words need not be in the lexicon. Consequently, the lexicon can also not be used to reduce the possible sequences of sub-word units. If certain words are not in the lexicon, it is still possible to restrict the number of acoustic transcriptions of an utterance. To this end preferably a statistical N-gram subword-string model 238 is used which provides the likelihood of the sequence of the last N subword units. For instance, a bigram could be used which specifies for each phoneme pair, the likelihood that those two phonemes follow each other. As such, based on general language characteristics the set of acoustic transcriptions of an utterance can be pruned, without the specific word (with its acoustic transcription) being in the lexicon. For non-full recognition, the unit matching subsystem 220 outputs the set 250 of acoustic transcriptions which correspond to an utterance.

For full recognition, it is preferred to also use a sentence level matching system 240 which, based on a language model (LM), places further constraints on the matching so that the paths investigated are those corresponding to word sequences which are proper sequences as specified by the language model. As such the language model provides the second term P(W) of equation (1). Combining the results of the acoustic model with the language model, results in an outcome of the unit matching subsystem 220 which is a recognised sentence (RS) 252. The language model used in pattern recognition may include syntactical and/or semantical constraints 242 of the language and the recognition task. A language model based on syntactical constraints is usually referred to as a grammar 244. The grammar 244 used by the language model provides the probability of a word sequence $W = w_1w_2w_3...w_q$, which in principle is given by:

$$P(W) = P(w_1)P(w_2|w_1).P(w_3|w_1w_2)...P(w_q|\ w_1w_2w_3...w_q).$$

Since in practice it is infeasible to reliably estimate the conditional word probabilities for all words and all sequence lengths in a given language, N-gram word models are widely used. In an N-gram model, the term $P(w_j| w_1w_2w_3...w_{j-1})$ is approximated by $P(w_j| w_{j-N+1}...w_{j-1})$. In practice, bigrams or trigrams are used. In a trigram, the term $P(w_j| w_1w_2w_3...w_{j-1})$ is approximated by $P(w_j| w_{j-2}w_{j-1})$.

The output of the speech recognition subsystem 40 (fully recognized speech and/or transcription sets) is fed to the dialogue management subsystem 50 of Fig.1. The dialogue manager 50 forms the core of the system 10 and contains application specific information. The dialogue manager 50 is programmed to determine in which information the user is interested. For a simple, fully controlled system the dialogue manager issues specific question statements, to which the user is expected to reply with only one utterance. In such a simple system, the dialogue manager can sometimes assume that the reply to a specific question represents the information the systems requires (without the system needing to recognize the utterance). In a more user-friendly system where the user may provide information in a free-speech style, the dialogue manager 50 scans the output of the speech recognizer in order to extract key-words or phrases which indicate which information the user wishes to obtain. The key words or phrases to be searched for may be stored in a storage 54, such as a harddisk, and be represented in the lexicon 234. This allows for full recognition of the keywords or phrases. Keywords, such as family names, which can not easily be fully recognized, need not be stored in this way. Contextual information which enables isolating an utterance representing a not-to-be-recognized keyword is preferably stored and recognized in full. For instance, for a directory assistance system, a user might say "I would like to have the phone number of Mr. Jones in London". This sequence of utterances contains two keywords, being the family name 'Jones' and the town 'London'. By including some or all other words of the sentence (and also many alternative ways of phrasing the question), it is possible to identify that indeed a family name and town are specified and extract the corresponding utterances. In this example, for a nation-wide or regional directory assistance systems it may be possible to accurately recognize the town. If so, it is preferred to add the town names (and acoustic transcriptions) to the lexicon for full recognition. The extracted information elements (recognized keywords/phrases or acoustic transcription sets), referred to as query items, are typically stored in main memory (not shown). Once all the essential items of a query have been recognized or identified and represented by a set of acoustic transcriptions, a search engine 59 is used to obtain the specified information from a storage 52. The storage is preferably based on a database, and the search engine 59 on search engines usually used for

searching databases. The storage is capable of storing a plurality of inquiry entries (e.g. directory entries), where each inquiry entry comprises a plurality of data fields, such as family name, street, town, and telephone number. It may be possible to specify a query in various ways. For instance, a directory query may be specified by a family or company name and a town. If the query performed by the search engine 59 results in too many hits (for instance only a frequently occurring family name and a name of a large town were given), additional information such as a street, house number or first name may be obtained in a further dialogue. A user may also have provided the additional information already in the main dialogue. Normally, the main dialogue is initiated by the dialogue manager 50 triggering the issuing of a predetermined welcome statement. If all the query items can be extracted from the initial user response, the dialogue may be completed after the first user response. Otherwise, one or more sub-dialogues may be started to determine the missing items. A sub-dialogue usually starts with a question statement. The question statement may be combined with a confirmation statement, to confirm already identified items. If a user rejects an item already identified, the rejected item may be re-determined, for instance from the rejection utterance, from less likely alternatives derived from previous utterances or by starting a new sub-dialogue. Particularly at the end of the dialogue, all the query items that have been identified may be confirmed by issuing a confirmation statement. The request/confirmation statements typically are formed from predetermined sentences/phrases, which may be stored in a background storage 56. A template sentence/phrase may be completed by the dialogue manager 50 filling in information believed to have been identified in utterances of the user. For instance, the dialogue manager could form a question/confirmation statement like "In which town lives Mr. Jones", where the name 'Jones' has been identified in a previous utterance. Also other variable information, such as current date or time, may be filled in.

The system 10 further comprises a speech generation subsystem 60. The speech generation subsystem 60 may receive the question/confirmation statements from the dialogue manager 50 in various forms, such as a (potentially prosodically enriched) textual form or as speech fragments. The speech generation subsystem 60 may be based on speech synthesis techniques capable of converting text-to-speech. The speech generation subsystem 60 may itself prosodically enrich the speech fragments or text in order to generate more naturally sounding speech. The enriched material is then transformed to speech output. Speech generation has been disclosed in various documents, such as IB 96/00770 (PHN 15408), IB 96/01448 (PHN 15641), US 5,479,564 (PHN 13801), all to the assignee of the present application. As described above, a question/confirmation statement may include information

identified by the dialogue manager 50. If such information was supplied in a fully recognized form (e.g. a textual representation), this information can be synthesized in a conventional manner, together with other elements of the statement. If the information was not fully recognized, the identified original input utterance may be used. Preferably, speech synthesis techniques are used to change the characteristics of the input utterance such that the user is not confronted with a sentence with certain system-specific voice characteristics (e.g. the voice of an actor) and one isolated utterance (his own) in between. Preferably, the prosody of the input utterance is changed to correspond to the prosody of the entire statement. Via the interface 70 the speech output is provided to the user at the speech output interconnection 80. Typically, a loudspeaker is connected to the interconnection 80 for reproducing the speech output. Preferably, the loudspeaker forms part of the telephone used for the speech input. The speech output interface 70 is usually combined with the speech input interface 30 in a known manner, for instance in the form of a modem.

According to the invention, the storage 52 stores for those elements which are not fully recognized a respective reference transcription set including at least two alternative acoustic transcriptions of data stored in a data field corresponding to the element. For instance, if only the family name is not fully recognized, than for each directory entry a set of acoustic transcriptions of the family name corresponding to that entry is stored. In principle, the family name need not be stored in a textual form. However, for most practical applications it will be useful to also store the family name as text, for instance for display purposes or for verifying whether the acoustic transcription set is accurate.

Fig. 4 illustrates a database structure for a directory assistance system. In the main table 400, each directory record (entry) is identified by an entry number (key) 402. In this example, each record provides information with respect to the following data fields: the family name 404, the initial(s) 406, the town 408, the street 410, the house number 412 and the telephone number 414. The fields family name 404, town 408 and street 410 are not searched for based on a textual representation but on an acoustic representation (set of acoustic transcriptions). This is illustrated further for the family name. In the family name field 404 a reference is stored to a record in a separate table 420 which stores the textual representation 422 of the family name. In this way the relatively long and frequently recurring family names need not be stored in full in the family name field 404. As will be clear from the following, also the reference set of acoustic transcriptions needs only be stored once. Each record of table 420 also comprises an identifier (key) 424 of the family name. In a third table 430, a reference acoustic transcription 432 is stored together with a reference 434 to the corresponding family

name. According to the invention several acoustic references, forming a set of reference acoustic references, are associated with one entry. To this end, all acoustic transcriptions in table 430 relating to the same record of table 420 refer back to the same identifier 424 of table 420. As an example, records 440, 442 and 444 of table 430 all relate to record 426 of table 420.

The reference acoustic transcription set may be obtained in any suitable way as for instance is known from training speech recognition systems. As an example, a selected group of people may be used to pronounce the item several times. Conventional techniques, such as the Viterbi algorithm, may be used to align the utterance with the states of the Hidden Markov Models and obtain a transcription of the utterance. As described earlier, such a reference transcription set is preferably pruned to a sufficiently small set of transcriptions to enable an adequately fast search of the storage.

In a preferred embodiment, the storage 52 stores for at least two data fields of the directory records respective associated transcription sets. Each transcription set includes at least two alternative acoustic transcriptions of data stored in the associated data field. In the example of Fig. 4, three data fields (family name, town, street) were associated with acoustic transcriptions sets. Conventionally, the search engine 59 can perform a query on two or more query fields in combination, where the query fields are specified by textual representations. In such a situation, a record is located by the query if all the query fields match the corresponding data fields of the record (a so-called AND operation). Such a query can be speeded up using conventional techniques, such as indexing on one or more of the query fields. Advantageously, the same combined query is performed where at least two of the query fields (and the associated data fields) are specified by sets of acoustic transcriptions. Where for one of the query field the number of hits may be very large, this will normally not be the case for a combined (AND) query. This avoids having to select between hits for individual query fields and as such no 'recognition' of the individual query fields is required. The query fields are only 'recognized' as belonging to the record(s) which was/were found to match the combined query.

In a preferred embodiment, a set of acoustic transcriptions is stored/processed as a graph. The graph representation may be used for the input transcription set(s) as well as the reference transcription set(s) stored in the storage 52. If a reference transcription set is represented as one graph, this makes table 430 of Fig. 4 obsolete. In the example of Fig. 4, the records 440, 442, and 444 can be replaced by one graph, which can be inserted in an additional field in table 420 for record 426. Fig. 5 illustrates a relatively simple graph. The nodes (501 to

515) represent an acoustic (sub-)word unit, such as an HMM state, like a phoneme. A path through the graph represents an acoustic transcription of the utterance. All paths through one graph are transcriptions of the same utterance. Normally, a likelihood is associated with the paths too. The graph of Fig. 5 consists of seven paths (transcriptions):

1. 501, 502, 503, 504
2. 501, 502, 503, 508, 509
3. 505, 502, 503, 504
4. 505, 502, 503, 509
5. 505, 506, 507, 508, 509
6. 510, 511, 512, 509
7. 510, 513, 514, 515

If only one of the involved transcription sets (the input transcription set or the reference transcription set) is represented as a graph, it is preferred that the search engine 59 matches the individual transcriptions of the set which is not represented as a graph against the graph. This may, for instance, be done by sequentially comparing the nodes of the individual transcription against the possible paths through the graph. Since, the paths of the graph have a common representation (they normally share nodes), this has the advantage of reducing the number of comparisons. For instance, if an individual path is compared with the graph of Fig. 5, a comparison with path 1 of the graph (as indicated above) may show a match of nodes 501 and 502, but a mismatch at node 503. This automatically implies that none of the other paths of the graph can be a match either: path 2 has coincided so far and therefore gives also a mismatch at node 503; the initial two nodes of paths 3 to 7 are different and cannot match the individual transcription. Similarly, if a mismatch only occurred at node 504, than only path 2 needs to be evaluated for a match of the remaining nodes 508 and 509.

Preferably, a match at node-level is not restricted to a binary yes-no decision. Instead, a measure of similarity between the nodes may be used. Such a measure may be based on a perceived acoustic similarity between sub-word units (nodes in the graph and transcription), such as phonemes. Preferably such a measure is combined with a likelihood measure of the paths to providing a likelihood of a match. A selection of most likely matches may be considered further. Instead of using a likelihood, also fuzzy logic rules may be used for determining whether a path and a graph match.

In the situation where both the input set of transcriptions and the reference set of transcriptions are represented by graph, it is preferred that the search engine 59 is

programmed to compare both graphs in one operation, benefiting of the node sharing in both graphs.

CLAIMS:

1. An automatic inquiry system including:

a storage for storing a plurality of inquiry entries, each inquiry entry comprising a plurality of data fields;

a dialogue engine for executing a machine-controlled human-machine dialogue to determine a plurality of pre-determined query fields;

a speech recognizer operative to represent an utterance as a corresponding input transcription set, which includes at least two alternative acoustic transcriptions of the corresponding utterance; a sequence of utterances specifying the respective query fields; and

a search engine for locating inquiry entries in the storage in dependence on the query fields;

characterized in that:

the storage is operative to store in association with each inquiry entry a respective reference transcription set including at least two alternative acoustic transcriptions of data stored in a data field of the inquiry entry;

the dialogue engine is operative to specify at least a first one of the query fields by an input transcription set of a selected one of the utterances; and

the search engine is operative to locate inquiry entries in the storage whose associated reference transcription set relates to the first one of the query fields.

2. A system as claimed in claim 1, characterized in that

the storage is operative to store in association with each inquiry entry at least a first and a second reference transcription set; for each inquiry entry the respective transcription sets being associated with a respective data field, and each transcription set including at least two alternative acoustic transcriptions of data stored in the associated data field;

the dialogue engine is operative to specify at least a first and second ones of the query fields by a respective input transcription set of respective selected ones of the utterances;

the search engine is operative to locate inquiry entries in the storage of which at least the associated first and second reference transcription set relate to the respective first and second ones of the query fields.

3. The system as claimed in claim 1, characterized in that:

in that the search engine is operative to determine that a reference transcription set relates to a query field specified by an input transcription set if at least one of the acoustic transcriptions of the reference transcription set is acoustically similar to at least one of the acoustic transcriptions of the input transcription set.

4. The system as claimed in 1, characterized in that the speech recognizer is operative to represent the input transcription set as an input graph; and in that the search engine is operative to determine that a reference transcription set relates to a query field specified by an input transcription set if at least one of the acoustic transcriptions of the reference transcription set is acoustically similar to at least one path through the input graph.

5. The system as claimed in claim 1, characterized in that the storage means is operative to store the respective reference sets of acoustic transcriptions as respective reference graphs; and in that the search engine is operative to determine that a reference transcription set relates to a query field specified by an input transcription set if at least one path through the reference graph representing the reference transcription set is acoustically similar to at least one acoustic transcription of the input transcription set.

6. The system as claimed in claim 1, characterized in that the speech recognizer is operative to represent the input transcription set as an input graph, referred to as input graph; in that the storage means is operative to store the respective reference sets of acoustic transcriptions as respective reference graphs; and in that the search engine is operative to determine that a reference transcription set relates to a query field specified by an input transcription set if at least one path through the reference graph representing the reference transcription set is acoustically similar to at least one path through the input graph.

7. The system as claimed in claim 1, characterized in that the dialogue engine comprises pruning means for pruning the acoustic transcriptions in the transcription set.

8. The system as claimed in claim 7, characterized in that the pruning means is operative to keep up to a predetermined number of acoustically most-likely acoustic transcriptions.

9. The system as claimed in claim 8, characterized in that the acoustic transcription represents a time-sequence of acoustic subword units, and in that the pruning means is operative to prune based on a statistical N-gram subword-string model.

10. The system as claimed in claim 7, characterized in that the pruning means is operative to prune an acoustic transcription by selecting a predetermined portion of the transcription.

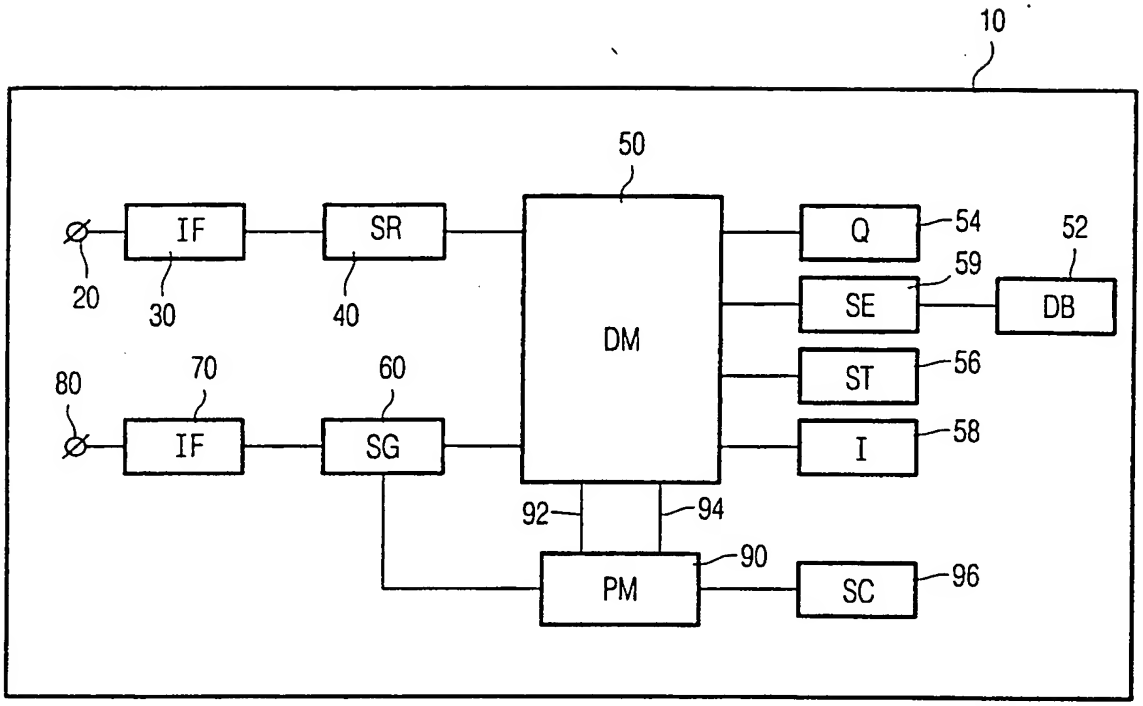

10
50
IF
SR
Q
54
52
20
30
40
59
SE
DB
DM
ST
56
80
70
60
IF
SG
I
58
92
94
90
PM
SC
96

FIG. 1

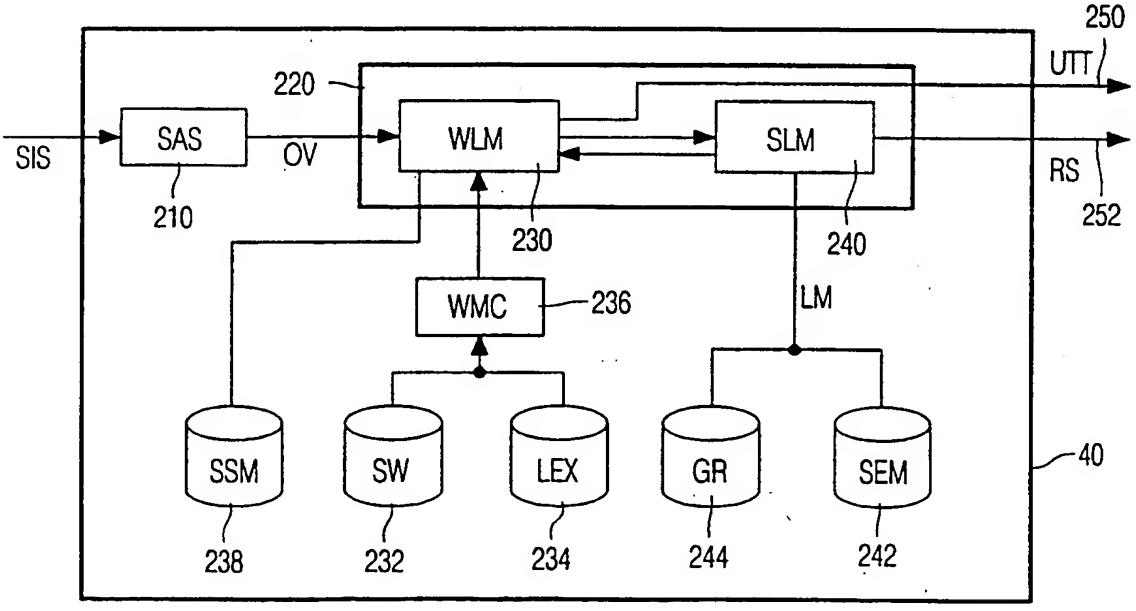

250
UTT
220
SIS
SAS
OV
WLM
SLM
RS
210
230
240
252
WMC
236
LM
SSM
SW
LEX
GR
SEM
40
238
232
234
244
242

FIG. 2

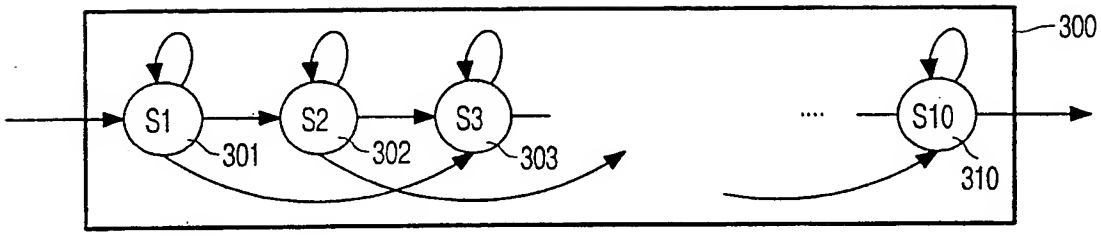

S1
S2
S3
S10
301
302
303
310
300

FIG. 3a

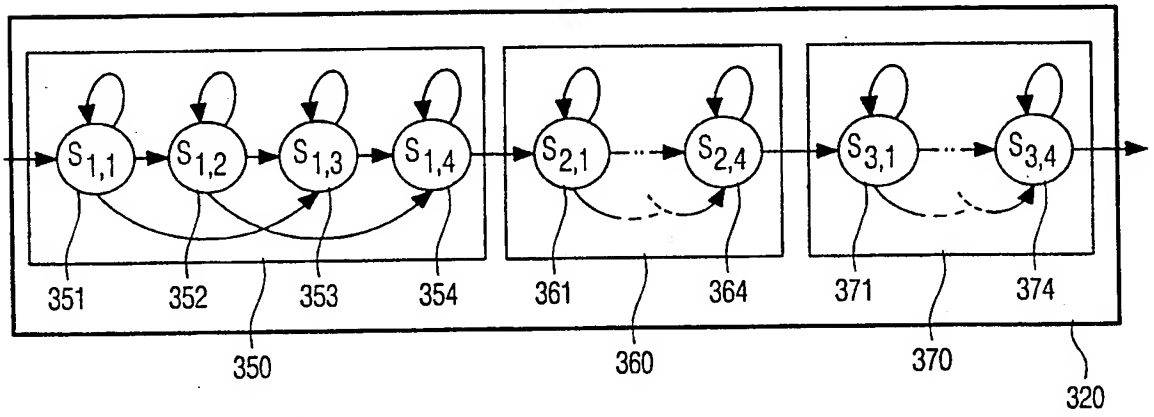

S1,1
S1,2
S1,3
S1,4
S2,1
S2,4
S3,1
S3,4
351
352
353
354
361
364
371
374
350
360
370
320

FIG. 3b

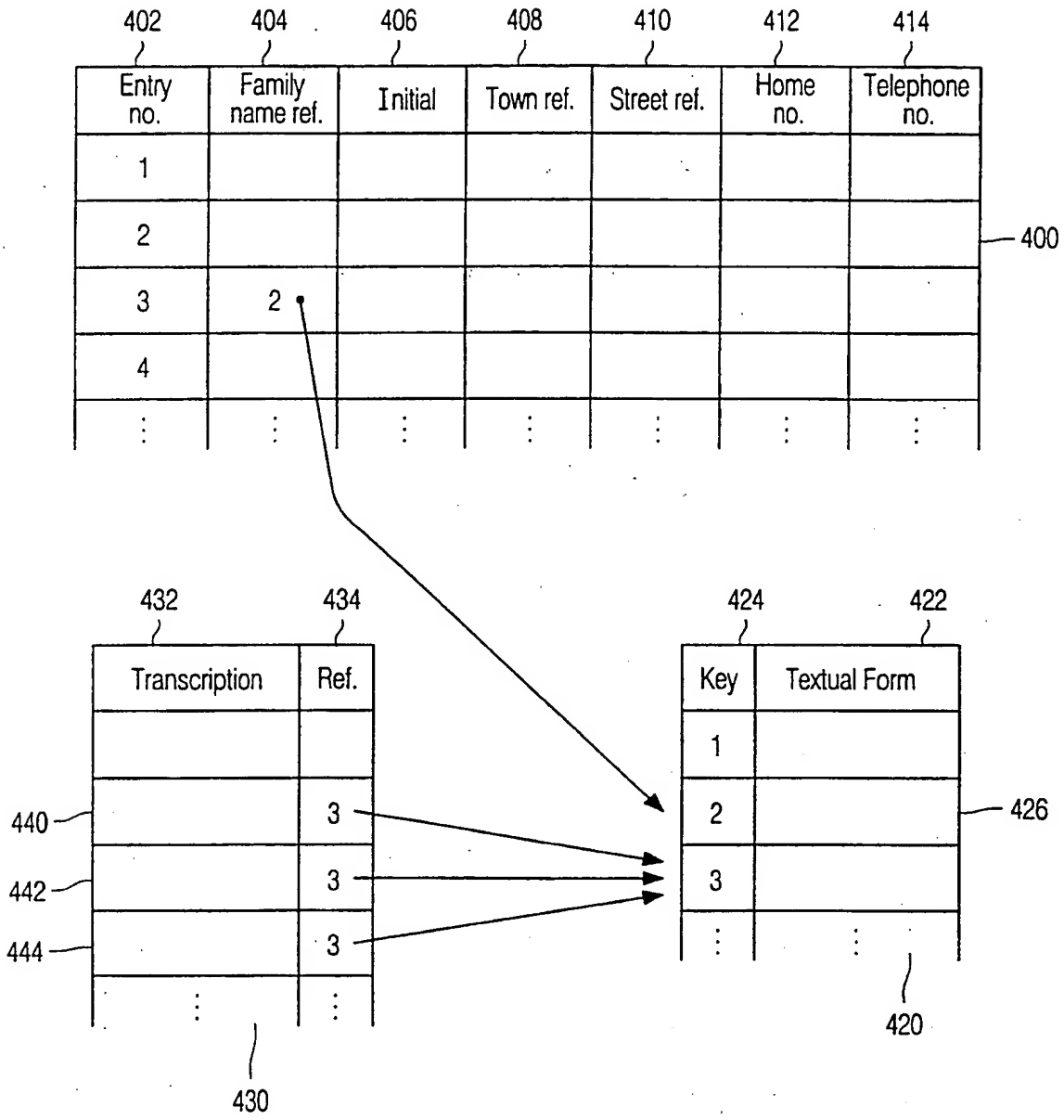

402
404
406
408
410
412
414
Entry no.
Family name ref.
Initial
Town ref.
Street ref.
Home no.
Telephone no.
1
2
3
4
2
400
432
434
424
422
Transcription
Ref.
Key
Textual Form
1
2
3
3
3
3
440
442
444
426
430
420

FIG. 4

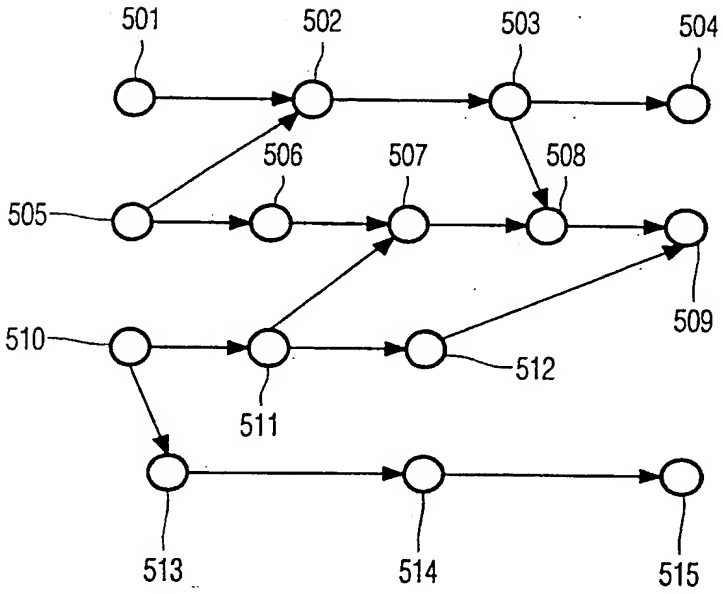


FIG. 5

INTERNATIONAL SEARCH REPORT

Inter. nal Application No
PCT/EP 99/09263

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L15/22

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 778 344 A (OLSEN PAUL A ET AL) 7 July 1998 (1998-07-07) column 1, line 29 -column 2, line 18 column 2, line 37 -column 4, line 44	1-3
Y		4-7
Y	US 5 581 655 A (COHEN MICHAEL H ET AL) 3 December 1996 (1996-12-03) column 2, line 24 -column 3, line 2 column 6, line 21 - line 53 figures 1,3	4-7
A	US 5 799 276 A (MALKOVSKY MIKHAIL ET AL) 25 August 1998 (1998-08-25) column 2, line 22 - line 37 column 11, line 42 -column 12, line 13	1-3,9

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 April 2000	26/04/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Ramos Sánchez, U

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inter. nal Application No
PCT/EP 99/09263

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	US 5 500 920 A (KUPIEC JULIAN M) 19 March 1996 (1996-03-19) column 9, line 38 - line 51 column 10, line 21 - line 28 column 12, line 62 -column 13, line 19	1-3
A	EP 0 564 166 A (AMERICAN TELEPHONE & TELEGRAPH) 6 October 1993 (1993-10-06) column 2, line 41 -column 4, line 4	1-3

1

INTERNATIONAL SEARCH REPORT

information on patent family members

Inter nal Application No
PCT/EP 99/09263

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5778344 A	07-07-1998	AU 707248 B	08-07-1999
		AU 3606897 A	22-08-1997
		CA 2244116 A	07-08-1997
		CN 1210643 A	10-03-1999
		EP 0878085 A	18-11-1998
		WO 9728634 A	07-08-1997
		NO 983501 A	30-09-1998
		NZ 326441 A	29-09-1999
US 5581655 A	03-12-1996	US 5268990 A	07-12-1993
		CA 2099978 A	01-08-1992
		EP 0573553 A	15-12-1993
		JP 6505349 T	16-06-1994
		WO 9214237 A	20-08-1992
US 5799276 A	25-08-1998	NONE	
US 5500920 A	19-03-1996	EP 0645757 A	29-03-1995
		JP 7175497 A	14-07-1995
EP 0564166 A	06-10-1993	CA 2088080 A,C	03-10-1993
		DE 69327188 D	13-01-2000
		FI 931471 A	03-10-1993
		JP 6012092 A	21-01-1994
		US 5329608 A	12-07-1994